# AUTHOR GUIDELINES FOR DCASE 2018 CHALLENGE TECHNICAL REPORT

Technical Report

*Chenchen Yu*

AI Lab, Lenovo Research
No.6 Shangdi West Road, Haidian District
Beijing, China
yucc1@lenovo.com

*Yu Hao*

AI Lab, Lenovo Research
No.6 Shangdi West Road, Haidian District
Beijing, China
haoyu5@lenovo.com

*Wenbo Yang*

AI Lab, Lenovo Research
No.6 Shangdi West Road, Haidian District
Beijing, China
yangwb3@lenovo.com

*Bo Fu*

AI Lab, Lenovo Research
No.6 Shangdi West Road, Haidian District
Beijing, China
fubo5@lenovo.com

**ABSTRACT**

For the task of Bird Audio Detection in the DCASE Challenge 2018[1], we present three approaches that all use convolutional neural networks on Mel-spectrogram. We obtained Area Under Curve (AUC) measure of 0.8610, 0.8548, 0.8464 on preview score which is calculated using approximate 1000 files randomly selected from the Chernobyl and warblrb10k data.

***Index Terms***— Mel spectrogram, Convolutional neural network, Bird audio detection

## 1. INTRODUCTION

Bird audio detection (BAD) is defined as identifying the presence or absence of bird call/tweet in a given audio recording. Detecting the presence of birdcalls in audio recordings can serve as a basic step for wildlife and biodiversity monitoring. BAD makes it possible to conduct work with large datasets (e.g. continuous 24h monitoring) by filtering data down to regions of interest. In order to advance the state of the art in automating this task, Stowell et al. [2] organized a Bird audio detection challenge as a subtask in the DCASE Challenge 2018. In this challenge, participants were asked to build algorithms to predict whether a given 10-second recording contains any type of bird vocalization or not.

The rest of the paper is organized as follows: the features are described in Section 2; the proposed CNN and its configuration for the BAD is explained and presented in Section 3; the description of ensemble used is in Section 4 and metrics and the results are report in Section 5.

## 2. FEATURE

In this work, we experiment three spectrograms including Mel-scaled log-magnitude spectrograms from 50Hz to 11kHz [3], Mel-scaled log-magnitude spectrograms from 2kHz to 11kHz and inverted Mel-scaled log-magnitude spectrograms from 50Hz to 11kHz, naming spectrogram 1-3.

We compute an STFT magnitude spectrogram with a window size of 1024 samples at 22.05 kHz sample rate with 70 frames per second. Then we apply a mel-scaled filter bank of n = 80 triangular filters from 50Hz to 11kHz(spectrogram 1) or from 2kHz to 11kHz(spectrogram 2) and scale magnitudes logarithmically. Alternative, we apply a inverted mel-scaled filter bank of n=80 triangular filters from 50Hz to 11kHz(spectrogram 3) and also scale magnitudes logarithmically.

The reason why we extract these features is that we observe that bird vocalizations are more concentrated in high frequency, especially frequency above 2k. The Mel-scaled filters usually emphasizes low frequencies, and we use inverted Mel-scaled filters to emphasize high frequencies.

| | |
|---|---|
| Input | 1×1000×80 |
| Conv(3×3) | 16×998×78 |
| Pool(3×3) | 16×332×26 |
| Conv(3×3) | 16×330×24 |
| Pool(3×3) | 16×110×8 |
| Conv(3×1) | 16×108×8 |
| Pool(3×1) | 16×36×8 |
| Conv(3×1) | 16×34×8 |
| Pool(3×1) | 16×11×8 |
| Dense | 256 |
| Dense | 32 |
| Dense | 1 |

Figure 1: The architecture of convolutional neural network

## 3. CONVOLUTIONAL NEURAL NETWORK

The network we used are all the same. The network also draw on the paper[3] and we change some parameter of it. This

convolutional neural network has a wide receptive field of 1000 frames(14s) and processes into a single binary output. There are four combinations of convolution and pooling condenses the input of 1000*80 into 16 feature maps of 11*8 units. Following three dense layers with 256, 32 and 1 unit(s) classify the condensed features. Each convolution and dense layer is followed by the leaky rectifier nonlinearity max(x, x/100) except for the sigmoid output layer. We can see the architecture on figure 1.
Training is done by stochastic gradient descent on mini-batches of 64, using the ADAM update rule[4] with an initial learning rate of 0.01 and 2000 epochs each. Features shorter than required are looped up to 1000 frames as need. All dataset are used to training each. As can be seen in table 1, we get three individual learner.

Table 1: Feature and model of the three individual learner.

| | Feature | Model |
|---|---|---|
| Learner 1 | Spectrogram 1 | CNN |
| Learner 2 | Spectrogram 2 | CNN |
| Learner 3 | Spectrogram 3 | CNN |

## 4. ENSEMBLE

Tree individual learners are trained and then predict the probability of bird presence each. Then, we ensemble these results. The submission 1 we present is the simple averaging the results of learner 1 and learner 2. The submission 2 is the simple averaging the results of learner 1, learner 2 and learner 3. The submission 3 is the simple averaging the results of learner 2 and learner 3.

## 5. EVALUATION AND RESULTS

The area under curve(AUC)[5] of the receiver operating characteristic(ROC) is used to evaluated the BAD system output. There are warblrb10k, Chernobyl and PolandNFC datasets. The preview score is calculated using approximate 1000 files randomly selected from the Chernobyl and warblrb10k data. The best preview score we got is 86.10.The preview scores of three submissions can be seen as table 2.

Table 2: The preview scores of AUC of three submissions each.

| preview score(%) | Submission 1 | Submission 2 | Submission 3 |
|---|---|---|---|
| AUC | 86.10 | 85.48 | 84.64 |

## 6. REFERENCES

[1] http://dcase.community/challenge2018/task-bird-audio-detection
[2] Stowell D, Stylianou Y, Wood M, et al. Automatic acoustic detection of birds through deep learning: the first Bird Audio Detection challenge[J]. arXiv preprint arXiv:1807.05812, 2018.
[3] Grill, T., & Schlüter, J. (2017, August). Two convolutional neural networks for bird detection in audio signals. In Signal Processing Conference (EUSIPCO), 2017 25th European (pp. 1764-1768). IEEE.
[4] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
[5] https://en.wikipedia.org/wiki/Receiver_operating_characteristic#Area under curve